Published in final edited form as:

Concepts Magn Reson Part A Bridg Educ Res. 2017 March ; 46A(2): . doi:10.1002/cmr.a.21437.

Covariance nuclear magnetic resonance methods for obtaining protein assignments and novel correlations

Aswani K. Kancherla and Dominique P. Frueh*

Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

Abstract

Protein NMR resonance assignment can be a tedious and error prone process, and it is often a limiting factor in biomolecular NMR studies. Challenges are exacerbated in larger proteins, disordered proteins, and often alpha-helical proteins, owing to an increase in spectral complexity and frequency degeneracies. Here, several multi-dimensional spectra must be inspected and compared in an iterative manner before resonances can be assigned with confidence. Over the last two decades, covariance NMR has evolved to become applicable to protein multi-dimensional spectra. The method, previously used to generate new correlations from spectra of small organic molecules, can now be used to recast assignment procedures as mathematical operations on NMR spectra. These operations result in multidimensional correlation maps combining all information from input spectra and providing direct correlations between moieties that would otherwise be compared indirectly through reporter nuclei. Thus, resonances of sequential residues can be identified and side-chain signals can be assigned by visual inspection of 4D arrays. This review highlights advances in covariance NMR that permitted to generate reliable 4D arrays and describes how these arrays can be obtained from conventional NMR spectra.

Keywords

Covariance NMR; Protein Sequence Specific Resonance Assignments; Methyl Resonance Assignments; Correlation Maps; Multidimensional NMR Spectra

INTRODUCTION:

Nuclear Magnetic Resonance (NMR) has become a mainstay in biomolecular studies, permitting measurement at or near physiological conditions. The technique is immensely versatile and allows the assessment of the chemical environments of atoms, the measurement of distances and bond orientations, and the characterization of dynamics at various timescales. Fluctuations of these parameters can be subsequently monitored during binding events,¹ biochemical reactions,^{2,3} or through cellular responses.^{4,5} Much of this versatility lies in the ability to obtain correlations between different nuclei within the molecule through a variety of correlation maps conveying site specific information at atomic resolution. For example, the environment of amide bonds can be monitored through HN-HSQC,⁶ and

*Correspondence to: dfrueh1@jhmi.edu.

NOESY⁷ spectra reveal protons near in space. However, before such information can be gleaned from NMR spectra, the identity of each correlation (or signal) in the spectrum must be paired with atoms in the molecule through resonance assignment. This process is the foundation of all molecular NMR studies and forms the cornerstone of meaningful spectral analysis, and much effort is dedicated to ensuring accurate assignments of NMR signals.

The assignment of protein NMR signals can be described as pairing so-called anchor correlations found in 2D HN-HSQC or 2D HC-HSQC according to spectroscopic features reporting on molecular topology. Pairing is usually conveyed via reporter nuclei that lead to common correlations in pairs of spectra for related anchors. For example, protein backbone resonances are assigned through a procedure called sequence-specific resonance assignment (Figure 1). Here, amide anchors belonging to sequential residues are identified through correlations with various carbon nuclei seen in complementary pairs of 3D spectra^{8–10} (Figure 1A). Each pair of spectra consists of a “sequential” spectrum, where amide protons and nitrogens are correlated to carbons of the previous residue, and an “intra” spectrum providing additional, stronger correlations with carbons of the same amino-acid. These spectra are obtained with triple-resonance experiments that exploit specific scalar couplings to funnel magnetization transfers (Figure 1B). Briefly, the sequential experiment HNCO⁹ utilizes the one bond scalar coupling $^1J_{\text{NC}}$ to convert the nitrogen single-quantum coherence of a residue into a carbonyl carbon (C') coherence for its preceding residue. Carbonyl carbon magnetization can be further transferred to C $^{\alpha}$ or C $^{\beta}$ carbon nuclei (in HN(CO)CA¹¹ and HN(COCA)CB, respectively), and each amide moiety can be correlated with carbon resonances (C', C $^{\alpha}$ or C $^{\beta}$) that belong exclusively to its preceding residue. Complementarily, the intra experiment HNCA¹⁰ utilizes the one-bond scalar coupling $^1J_{\text{NC}\alpha}$ for transferring nitrogen magnetization to the C $^{\alpha}$ of the same residue; in HN(CA)CB¹² and HN(CA)CO¹³ carbon-to-carbon transfers allow for correlating C $^{\beta}$ and C' with amide moieties of the same residue. Because the magnitudes of two-bond scalar couplings $^2J_{\text{NC}\alpha}$ between sequential residues are similar to those of $^1J_{\text{NC}\alpha}$, low intensity correlations with carbons from previous residues often appear in HNCA, HN(CA)CB, and HN(CA)CO. All resulting 3D spectra span dimensions reporting amide proton and nitrogen chemical shifts as well as carbon chemical shifts. As this review focuses on extracting information from these spectra rather than the experiments that provide them, we introduce a notation highlighting their dimensions and characteristic correlations. For example, an HN(CO)CA leads to a spectrum described with $[\text{H}_s][\text{N}_s][\text{CA}_s]$, where s denotes a sequential experiment, and its signals provide correlations $(\text{H}_r, \text{N}_r, \text{C}^{\alpha}_{r-1})$, where r denotes the residue number. The square brackets identify dimensions, e.g. the dimension along proton frequencies is $[\text{H}_s]$, and the regular brackets define a correlation centered on a signal observed in the spectrum. We will see below that such a notation allows for distinguishing an arbitrary data point in a spectrum, for example identified with $[\text{H}][\text{N}][\text{CA}](i,j,k)$, from a correlation between signals of various nuclei such as $(\text{H}_r, \text{N}_r, \text{C}^{\alpha}_{r-1})$. The three pairs of spectra used for backbone resonance assignments are $[\text{H}_s][\text{N}_s][\text{CA}_s]$ and $[\text{H}][\text{N}][\text{CA}]; [\text{H}_s][\text{N}_s][\text{C}'_s]$ and $[\text{H}][\text{N}][\text{C}'];$ and $[\text{H}_s][\text{N}_s][\text{CB}_s]$ and $[\text{H}][\text{N}][\text{CB}]$. The notation also clearly highlights that amide anchors are common to all spectra, be they in $[\text{H}][\text{N}]$ or $[\text{H}_s][\text{N}_s]$ dimensions, but carbon reporter nuclei differ between pairs.

By comparing carbon correlations present in “intra” and “sequential” spectra, one can in principle identify amide groups of sequential residues. For example, while HNCA provides a

correlation of the form (H_r, N_r, C_r^α) , HN(CO)CA provides $(H_{r+1}, N_{r+1}, C_r^\alpha)$. In these spectra, every correlation seen in a 2D HN-HSQC spectrum is found in an $[H][N]$ plane, now at the frequency of the relevant carbon. Thus, to identify sequential residues, one could compare $[H][N]$ planes of HNCA with $[H_s][N_s]$ planes of HN(CO)CA, both at the same coordinate C_r^α .¹⁴ However, many (H,N) correlations are observed in each $[H][N]$ plane because of near degeneracy of C^α frequencies together with line broadening along $[CA]$ and $[CA_s]$. To overcome this issue, $[H][N]$ planes should also be compared at C_r^β and C'_r , which each hopefully displays different (H,N) correlations except for (H_r, N_r) and (H_{r+1}, N_{r+1}) that should appear at C_r^α , C_r^β , and C'_r (Figure 1A). In practice, such a procedure is too time consuming and prone to human error because the prevalence and severity of signal overlap often impedes unambiguous visual assessment. Instead, each 3D spectrum is regarded as a collection of 1D ^{13}C spectra at positions defined by 1H and ^{15}N chemical shifts. That is, $[CA]$ and $[CA_s]$ dimensions are compared for all (H,N) correlations and likewise for $[C']$ and $[C'_s]$ and $[CB]$ and $[CB_s]$ dimensions, respectively. In practice, the procedure necessitates peak picking and help from NMR spectral analysis software such as CARA,¹⁵ Sparky,¹⁶ or CcpNmr Analysis¹⁷ to extract strips along $[H][C]$ dimensions. The principal steps are summarized as follow:

1. Pick all peaks present in the 2D HN-HSQC anchor spectrum and provide them with an arbitrary label. In larger proteins, (H,N,C') correlations are picked in an HNCO to overcome overlap.
2. Visualize ^{13}C strips from the 3D spectra for each correlation present in the anchor spectrum and pick carbon signals. The procedure identifies frequencies of (H,N,C) correlations.
3. Compare strips of (H,N) correlations to identify pairs of correlations that are sequentially related. This step usually relies on sorting strips according to differences in carbon frequencies defined in step 2 for intra and sequential spectra (Figure 1C).
4. Residue types (alanines, glycines, etc.) are tentatively assessed through more or less characteristic C^α , C^β , and C' chemical shifts.
5. Strips of hypothetical sequential residues are displayed side-by-side to obtain a sequence of residue types which can be compared with that of the protein primary structure. The signals are assigned when the residues inferred from the strips unambiguously and uniquely match a region of the primary structure.

This assignment process has the advantage of being relatively straightforward and amenable to semi-automation for proteins with favorable spectroscopic properties. However, increased chemical shift degeneracy and spectral crowding hinder peak picking in steps 1 and 2 and make the procedure error prone, a problem which is common in alpha-helical proteins, unfolded proteins, and in larger proteins where increased linewidths further exacerbate ambiguities. Here, increased chemical shift degeneracy along carbon dimensions dictates that multiple carbon correlations should be tested simultaneously to identify sequential residues (Figure 1C). Unfortunately, overlapping (H,N) correlations are frequent, which leads to multiple signals along $[C]$ dimensions in the same strip (Figure 1C, 48/52). For

successful assignment, the signals must be paired with the right combination, and a series of arbitrary and error-prone combinations of carbon frequencies must be considered in the process. The correct combination can only be identified by trial and error, which can consume significant amount of time. This is illustrated in Figure 1C, where strips matching either of the target carbons must be displayed for all spectra and compared. Note that for clarity only a few representative strips are displayed in this figure. Further, signals of HN(COCA)CB are often weak, and the absence of a matching strip in $[H_s][N_s][CB_s]$ cannot be used as criterion. In short, the successful assignment of resonances of larger proteins would require simultaneous analysis of all spectra, but the procedure above relies on iterative peak picking steps that must be performed at the onset, before information can be combined. We will show that by translating the concepts described in this section into mathematical operations, it is possible to combine all information carried by the spectra into a single correlation map and overcome many challenges when assigning crowded spectra.

Assignments of side-chain resonances are usually obtained by pairing their signals with those of assigned amide resonances. In smaller proteins, the experiments H(CCO)NH and (H)C(CO)NH¹⁸ provide direct correlations between assigned amide anchors and side-chain proton and carbon resonances, respectively. An HC(C)H TOCSY¹⁹ is used to pair protons with carbon resonances and hence to assign (H,C) anchors in HC-HSQC. However, in larger proteins, adverse transverse relaxation deteriorates the quality of H(CCO)NH and (H)C(CO)NH spectra, and methyl resonances are assigned indirectly in a manner reminiscent of sequence specific assignments. Thus, methyl groups are correlated with C^α and C^β for comparison with existing assignments obtained with HNCA and HN(CA)CB.²⁰ That is, [CA] and [CB] dimensions of $[H][N][CA]$ and $[H][N][CB]$ spectra are compared with those of $[HM][CM][CACB]$ spectra, in which M denotes dimensions specific to methyl groups (e.g. γ for valines, δ , for leucines or isoleucines) and [CACB] is a dimension featuring C^α and C^β signals. In closing, carrying out resonance assignments is a time-consuming task, and many a time it is the bottle-neck for structural and functional studies of larger proteins by solution NMR.

Covariance NMR methods can be used to overcome some of the challenges in the assignment process described above. Here, NMR spectra are treated as matrices or multi-dimensional arrays, and assignment processes are formulated as mathematical operations. Most importantly, such covariance methods bypass the need for *initial* peak picking and hence avoid all the errors associated with this step. In essence, covariance NMR is used to generate new correlations from a combination of spectra. However, such applications had initially been limited to analysis of sparse spectra, e.g. small molecules and metabolites, because false correlations (or artifacts) abound in covariance correlation maps of crowded spectra such as those of proteins. Recent advances in covariance processing methods significantly alleviate such artifacts and thus helped extend covariance NMR applications to analysis of protein spectra, including larger proteins. Theoretical aspects of covariance NMR and its applications to analysis of small molecule spectra have recently been reviewed by Snyder and Bruschweiler.²¹ Here, we focus on concepts necessary for understanding applications for protein resonance assignments.

The next two sections provide a brief historical introduction to covariance methods and gradually introduce concepts necessary for applications in protein resonance assignments. We next describe artifacts present in covariance maps and solutions to reduce these artifacts. Finally, we reformulate protein resonance assignments through covariance processing and describe additional operations introduced to remove artifacts. With these advances, the method is amenable for analyzing crowded spectra, as exemplified with data of a nonribosomal peptide synthetase cyclization domain (52 kDa).

NMR SPECTRA AS MATRICES AND NUMERICAL ARRAYS

In order to understand covariance NMR processing, NMR spectra must be treated as matrices or arrays of numbers. For instance, a one-dimensional NMR spectrum $s(\nu)$ with M data points can be represented by a one-dimensional array of numbers $s(m)$, that is a column vector of size $(M \times 1)$, as shown in Figure 2A. The value of each element represents the intensity of the signal in the spectrum at the corresponding data point, including noise in absence of signal. The value of the frequency (or similarly the chemical shift) associated with each data point is conveyed by its index and can be obtained with:

$$\nu(m) = \nu_0 - SW/2 + m(SW/M), \quad (1)$$

in which $\nu(m)$ is the frequency at index m , ν_0 is the carrier frequency, and SW is the spectral width, all in Hz. Similarly, a two-dimensional spectrum with M and N data points in the detected and indirectly recorded dimensions, respectively, can be represented as a two-dimensional array of numbers, that is a matrix of dimensions $(M \times N)$ as shown in Figure 2B. Finally, 3D and higher order spectra can be represented as higher order multi-dimensional arrays of numbers depicting signal amplitudes at frequencies defined by the indices of each element. These arrays can themselves be described through vectors and matrices, a fact we will rely upon when describing mathematical operations used to generate or optimize novel spectra. For example, a 2D array may sometimes best be seen as a column of row vectors, or a 4D array may be considered as a 2D array of matrices. With such representations, it is possible to predict the outcome of matrix algebra on NMR spectra of various dimensions.

Covariance NMR:

In its most basic definition, covariance NMR is used to identify correlations within a set of spectra, originally one-dimensional. The first application of covariance NMR consisted in increasing the resolution of 2D symmetric spectra, such as HH-NOESY or TOCSY spectra.²² Here, a 2D NMR spectrum with elements $S(m,n)$, M data points in the detected dimension, and N points in the indirect dimension can be regarded as a series of 1D spectra in two different ways. First, the data matrix can be considered as an array of N row vectors of size M . This representation is traditionally used to describe that a series of 1D spectra with amplitudes encoding a second dimension leads to a correlation map upon Fourier transform of the second dimension. Alternatively, the spectrum can be regarded as a series of M column vectors $S(n)$ of size N each. With this representation it is possible to retrieve

correlations between these M vectors by analyzing their pairwise covariance through a symmetric ($M \times M$) covariance matrix $C_{M,M}$ whose elements are defined by:²²

$$C_{M,M}(i, j) = \frac{1}{N-1} \sum_{k=1}^N (S(k, i) - \langle S(i) \rangle)(S(k, j) - \langle S(j) \rangle), \quad (2)$$

where $\langle S(i) \rangle$ is the average intensity over the spectrum at index i , given by:

$$\langle S(i) \rangle = \frac{1}{N} \sum_{k=1}^N S(k, i). \quad (3)$$

In the covariance matrix, each element $C_{M,M}(i, j)$ depicts the amplitude of the covariance between the column vectors $S(i)$ and $S(j)$ with resonance frequencies $\nu(i)$ and $\nu(j)$ given by equation (1). If the vectors are correlated, that is if they both report on a common frequency in the indirect dimension, a correlation will emerge at (i, j) . In absence of correlation, the amplitude will vanish or be reduced at (i, j) . Equation (2) can be applied equivalently to vectors where the indirect dimension is in the time domain or in the frequency domain. In the frequency domain, the procedure simply correlates $S(i)$ and $S(j)$ when they possess common signals. Thus, for symmetric spectra such as NOESY or TOCSY, all cross-peaks in the spectrum $S(m, n)$ obtained by Fourier transform of the indirect dimension will appear as correlations in the symmetric matrix $C_{M,M}(i, j)$, with the advantage that both dimensions now benefit from the high resolution of the detected dimension.^{22,23}

Simple inspection of equation (2) reveals that it closely resembles matrix multiplication, save for subtracting the average spectrum of each vector $S(i)$ and $S(j)$. Indeed, it was rapidly demonstrated that multiplying a symmetric spectrum by its transpose produces a square matrix with elements faithfully describing the spectrum upon matrix square-rooting²³ (see more about the role of square-rooting in artifact suppression below). This spectrum is related to the covariance matrix by:

$$S_{M,M}(i, j) = \sqrt{2} C_{M,M}^{\frac{1}{2}}(i, j), \quad (4)$$

or in matrix form

$$S_{M,M} = (S^T S)^{1/2} = \sqrt{2} C_{M,M}^{\frac{1}{2}}. \quad (5)$$

In effect, it is the left-hand side equality of (5) that is most commonly employed, although the name covariance NMR has remained in use. As covariance is recast as a matrix product, the considerations we made about matrices can be reformulated in light of NMR spectroscopy. Each element of the covariance matrix corresponds to the scalar product of row and column vectors in the spectrum and its transpose, respectively. As they represent

slices in the spectrum, the value of this scalar is large and positive if both vectors feature signals at common indices and vanishes when they don't. This formulation will prove useful when investigating the source of common artifacts in reconstructed spectra and in understanding how they can be minimized.

Covariance NMR for obtaining new correlations:

In 1993, Munk and Co-workers demonstrated that new correlations could be obtained by multiplying different NMR spectra²⁴. HH-COSY was multiplied with HC-HSQC both on the left- and right-hand sides to generate a CC-COSY spectrum. Similarly, the formalism described in the previous section was extended so as to produce correlations absent from a given NMR spectrum.²⁵ Here, the advantage was that artifacts could be reduced through matrix square-rooting (see next section). The first example of such an “indirect covariance” was demonstrated by Zhang and Bruschweiler in 2004²⁵ using a 2D HC-HSQC-TOCSY that provides a spectrum $[H_t][C_t]$, in which TOCSY patterns (t) appear for every proton and carbon signals belonging to the same scalar coupling network (Figure 3A). Here, covariance of the detected proton dimension for all carbon indices provides a $[C_t][C_t]$ correlation map, as easily revealed through matrix formulation (Figure 3B):

$$[C_t][C_t](i, j) = \sum_{k=1}^K [H_t][C_t](k, i) \cdot [H_t][C_t](k, j). \quad (6)$$

It is worth emphasizing that the subsumed dimension is the detected proton dimension, with highest resolution. This indirect covariance correlation map corresponds to a CC-TOCSY spectrum (Figure 3B), which however originates from a proton excited, proton detected experiment and hence benefits from much higher sensitivity when compared to a ¹³C excited, ¹³C detected CC-TOCSY²⁶ spectrum. This application exemplified how covariance could be used to provide spectra with otherwise prohibitive acquisition times. However, this example is conceptually identical to that of the direct covariance approach described above as the correlations, although new in nature, are nevertheless obtained by covarying a spectrum with itself, and the resulting indirect covariance correlation map corresponds to a symmetric square matrix.

Soon, Blinov et. al.,²⁷ showed that novel correlations can also be generated from two different spectra using covariance. Such an application was illustrated by covarying proton dimensions of HC-HSQC and HMBC²⁸ spectra, of the form $[H][C]$ and $[H][C_{mb}]$, where “mb” stands for multiple bonds, to obtain long-range carbon-carbon connectivities:

$$[C][C_{mb}] = ([H][C])^T \cdot [H][C_{mb}]. \quad (7)$$

This type of indirect covariance exemplified that any two different spectra could be covaried provided they share common correlations in one dimension. For example, covariance of an HC-HMBC spectrum $[H][C_{mb}]$ with an HH-TOCSY $[H_t][H_t]$ spectrum results in a $[H_t][C_{mb}]$ spectrum as obtained from HC-HMBC-TOCSY,²⁹ whereas covariance of HC-HSQC and HH-COSY provides an artificial spectrum comparable to that of HC-HSQC-COSY³⁰.

Similarly, HC-HMBC $[H][C_{mb}]$ and HN-HSQC $[H][N]$ spectra lead to a $[C_{mb}][N]$ correlation map,^{31,32} while HC-HMBC $[H][C_{mb}]$ and HN-IMPEACH-MBC $[H_{mb}][N]$ provided an increased number of long range correlations in $[C_{mb}][N]$.³³ The concept was also incorporated into the powerful hyperdimensional NMR formalism, where it is paired with projection reconstruction to provide access to higher dimensionality spectra.³⁴ In all these applications, the covaried dimensions are subsumed in contrast to a derivatized approach named cross-spectra in which matrix elements are multiplied but indexed instead of being subsumed.³⁵ Finally, we note that 2D covariance maps obtained through indirect covariance may lead to asymmetric, rectangular matrices, a property which required additional theoretical efforts for artifact suppression (next section).

Artifacts in covariance spectra and solutions:

False correlations, or artifacts, have limited general applications of covariance NMR, in particular for protein spectra. One class of artifacts results from relay effects mediated by accidental frequency degeneracies. Such artifacts occur in the example described above when a $[C_l][C_l]$ map is obtained through an HSQC-TOCSY. Here, true (C,C) correlations, i.e. reflecting molecular connectivities, are obtained when both carbon indices carry the signals of protons within the same scalar coupling network spanned by the TOCSY sequence. However, false (C,C) correlations will emerge when carbon indices reporting on different networks each feature one or more protons with accidental degeneracies (Figure 4). In other words, two carbons displaying different patterns along the $[H]$ dimension of HSQC-TOCSY but featuring one or more signals at the same 1H frequency will lead to a (C,C) correlation regardless of whether they belong to the same scalar coupling network. Blinov et al.³⁶ demonstrated that indirect, unsymmetrical covariance processing of spectra of a modified HSQC-TOCSY experiment permitted removal of such artifacts. A more general method to reduce these artifacts through square-rooting of the covariance matrix had been designed by Zhang and Bruschweiler,²⁵ and matrix square-rooting was facilitated (and computationally optimized) through Singular Value Decomposition (SVD, see below).³⁷ However, matrix square-rooting was a priori not applicable for rectangular, asymmetric matrices obtained when two different spectra are covaried. In 2009, Snyder and Bruschweiler²⁹ proposed an elegant Generalized Indirect Covariance (GIC) NMR formalism to embed processing of asymmetric, rectangular covariance matrices within the context of the regular covariance transform.

GIC enables both (i) removal of artifacts in covariance spectra with different sizes along their dimensions and (ii) simultaneous processing of more than two spectra. We begin our discussion by considering only two spectra that result in an asymmetric covariance map and move on to situations involving an arbitrary number of spectra. Let X_1 and X_2 be two matrices that represent 2D spectra with sizes $N_{1,1} \times N_2$ and $N_{1,2} \times N_2$, respectively, to be covaried along the second dimension of size N_2 and with $N_{1,1} \neq N_{1,2}$. The covariance matrix C is given by:

$$C = X_1 \cdot X_2^T. \quad (8)$$

To calculate a square-root of this non-symmetric matrix C , a larger square matrix is generated using all four possible covariances involving the two matrices X_1 and X_2 : two self-covariance matrices involving a matrix multiplied by its own transpose and two cross-covariance matrices. The first cross-covariance matrix corresponds to equation (8) and the second is obtained with permutation of the indices 1 and 2 in equation (8). The matrix encompassing all covariance matrices was defined as a GIC matrix

$$C_{GIC(2)} = \begin{bmatrix} X_1 X_1^T & X_1 X_2^T \\ X_2 X_1^T & X_2 X_2^T \end{bmatrix}, \quad (9)$$

where the subscript (2) denotes the number of spectra involved. The covariance matrix of interest $X_1 X_2^T$ is present as an off-diagonal submatrix within the larger generalized covariance matrix $C_{GIC(2)}$. The matrix $C_{GIC(2)}$ by itself is a square matrix and hence amenable to matrix square-rooting and related matrix manipulations. A square-root of the covariance matrix of interest, $C = X_1 \cdot X_2^T$, can be obtained by calculating a square-root of $C_{GIC(2)}$ and then extracting the submatrix from the region corresponding to C . Before describing these operations, we first show that the procedure can be generalized in a compact form for an arbitrary number of matrices by defining an array of matrices, S :

$$S = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}, \quad (10)$$

where $X_1, X_2, X_3, \dots, X_n$ are n 2D spectra of dimensions $N_{Ii} \times N_2$ ($i = 1, 2, 3, \dots, n$). The generalized covariance matrix $C_{GIC(n)}$ is then defined as:

$$C_{GIC(n)} = S \cdot S^T = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix} \cdot \begin{bmatrix} X_1^T & \cdot & \cdot & \cdot & X_n^T \end{bmatrix}, \quad (11)$$

Or

$$C_{GIC(n)} = \begin{bmatrix} X_1 X_1^T & \dots & X_1 X_n^T \\ \vdots & \ddots & \vdots \\ X_n X_1^T & \dots & X_n X_n^T \end{bmatrix}. \quad (12)$$

Here, the GIC matrix $C_{GIC(n)}$ is a square matrix with dimensions $(N \times N)$, where $N = \sum_{i=1}^n N_{1i}$, which is amenable to matrix square-rooting. Again, operations on the desired covariance matrix, e.g. $X_1 \cdot X_2^T$ or $X_1 \cdot X_n^T$, are obtained by operating on $C_{GIC(n)}$ and extracting the relevant quadrant a posteriori. Square-roots, or any arbitrary powers, of $C_{GIC(n)}$ are obtained via SVD of the spectral array S defined in equation (10):

$$S = U \cdot D \cdot V^T, \quad (13)$$

in which both U and V are orthogonal such that $U^T \cdot U = U \cdot U^T = I$ and $V^T \cdot V = V \cdot V^T = I$, where I is the identity matrix. Substituting S from equation (13) in equation (11), and exploiting the orthogonality of V , we get:

$$C_{GIC(n)} = (U \cdot D \cdot V^T) \cdot (V \cdot D \cdot U^T) = (U \cdot D \cdot V^T \cdot V \cdot D \cdot U^T) = (U \cdot D^2 \cdot U^T). \quad (14)$$

The covariance matrix raised to the power λ is simply given by

$$C_{GIC(n)}^\lambda = U \cdot D^{2\lambda} \cdot U^T, \quad (15)$$

and the square root of the GIC matrix is

$$C_{GIC(n)}^{0.5} = U \cdot D \cdot U^T. \quad (16)$$

The λ^{th} power of the non-symmetric covariance matrix of interest is obtained by extracting the corresponding quadrant from $C_{GIC(n)}^\lambda$. Importantly, many spectra can be rooted from a single SVD, which results in considerable savings in computing time. This gain in time is exploited when calculating maps from 3D spectra, as each plane can be considered as an individual 2D spectrum to be stacked as in equation (10).³⁸ For example, to calculate covariance maps for sequence specific resonance assignments (next section), different planes of the intra and sequential 3D spectra are treated as series of 2D spectra stored in an array similar to S in equation (10). All covariance matrices, raised to the desired power, are then calculated in a single step to produce a 4D array.

Figure 4 illustrates the advantage of matrix square-rooting in covariance NMR.²⁹ In this simulation, 2D HMBC (Figure 4A) and TOCSY (Figure 4B) spectra are covaried along the proton dimension to obtain a 2D HMBC-TOCSY spectrum. We consider two separate hypothetical spin systems $\{C_xH_x, C_yH_y, C_zH_z\}$ and $\{C_uH_u, C_vH_v\}$ (Figure 4E), in which two protons, H_y and H_u , feature degenerate frequencies. This degeneracy leads to several relay artifacts in the covariance spectrum of power 1 (Figure 4C, in red) in addition to the patterns reflecting the true scalar coupling networks (green, and blue). In contrast, the majority of these artifacts are suppressed in a correlation map of power 0.5, and the surviving artifacts are greatly reduced in intensity (Figure 4D). Hence, the methodology described in this section has been implemented within the framework of protein NMR resonance assignment. We will see in the following section that further developments were necessary to overcome challenges in protein spectra.

COVARIANCE NMR FOR PROTEIN RESONANCE ASSIGNMENT

Sequence specific backbone resonance assignment

The complex assignment procedures described in the introduction and in Figure 1 can gradually be reformulated to build a mathematical equivalence. For simplicity, this section is described using C^α signals in HNCA and HN(CO)CA but the same considerations apply to C^β and C' carbon signals in their respective experiments. For all amide anchors in HNCA each featuring an intra-residue C^α signal, we seek to identify an amide anchor in HN(CO)CA with a matching sequential C^α signal. Using our notation to highlight dimensions, we can state that we compare $[CA_s]$ and $[CA]$ dimensions to identify all (H_r, N_r, N_{r+1}) correlations of the spectrum $[H_s][N_s][CA_s]$ corresponding each to an (H_r, N_r) correlation in $[H][N][CA]$. Finally, the objective can be described as identifying the indexes l and m in an array $[H_s][N_s][CA_s](l, m, k)$ that correlate with indexes i and j in $[H][N][CA](i, j, k)$ based on signals in $[CA]$ and $[CA_s]$. Building upon the concepts introduced in the previous sections, the comparison of $[CA_s]$ and $[CA]$ dimensions corresponds to a covariance analysis, whereas identifying all correlations requires spanning all indexes in proton and nitrogen dimensions of both $[H_s][N_s][CA_s]$ and $[H][N][CA]$. That is, we build a covariance array of the form:

$$[H][N][H_s][N_s](i, j, l, m) = \sum_{k=1}^K [H][N][CA](i, j, k) \cdot [H_s][N_s][CA_s](l, m, k). \quad (17)$$

To facilitate comparison with the descriptions provided above, we recast this array as a series of covariance matrices obtained from $[H][CA]$ and $[H_s][CA_s]$ planes spanning $[N]$ and $[N_s]$ dimensions (Figure 5A, B):

$$[H][H_s](i, l) = \sum_{k=1}^K [H][CA](i, k) \cdot [H_s][CA_s](l, k). \quad (18)$$

This notation highlights that strip comparisons have now been translated into direct correlations between protons of sequential residues. Hence, arrays provided by equation (17) are called Covariance Sequential Correlation Map (COSCOM).^{38,39} Here, the amplitude of

$[H][H_s](i,l)$ is only significant when amplitudes along $[CA_s]$ match those along $[CA]$, i.e. when a signal coincides in both spectra. Unfortunately, for such a correlation to appear, it is only sufficient for the amplitudes of the signals to match, and it is not necessary for them to have the same maxima. In proteins, many residues possess carbons with similar frequencies, and a multitude of correlations appear in COSCOMs even though a strip comparison would reveal that many do not belong to sequential residues. Such an artifact correlation is depicted conceptually in Figure 5E. This issue is best analyzed by further reformulating the matrix product in equation (18) as a series of scalar products along the $[CA]$ and $[CA_s]$ dimensions. Here, each element of the covariance matrix $[H][H_s](i,l)$ has an amplitude H_{il} defined with

$$H_{il} = \sum_{k=1}^K [CA](k) \cdot [CA_s](k) = \sum_{k=1}^K \vec{C}(k) \vec{C}_s(k), \quad (19)$$

in which the vectors $\vec{C}$ and $\vec{C}_s$ represent the one-dimensional arrays $[CA](k)$ and $[CA_s](k)$, respectively. The scalar product in equation (19) can be further described as the sum of the elements of a vector $\vec{v}$ obtained through the element-wise product of $\vec{C}$ and $\vec{C}_s$

$$\sum_{k=1}^K \vec{C}(k) \vec{C}_s(k) = \sum_{k=1}^K \vec{v}(k), \quad (20a)$$

$$\vec{v} = \vec{C} \odot \vec{C}_s, \quad (20b)$$

where $\odot$ denotes an element-wise multiplication. Equations (20) demonstrate that the amplitude of a covariance matrix can be assessed by inspecting the elements of vectors $\vec{v}$. Figures 5C, D, F and G illustrate graphically such an analysis to compare correlations for true sequential residues with those stemming from signals with similar but different carbon frequencies. Here, the vector $\vec{C}$ corresponds to a slice along the $[CA]$ dimension of HNCA, the vector $\vec{C}_{s*}$ describes a slice along $[CA_s]$ in HN(CO)CA for a true sequential residue, and $\vec{C}_{sx}$ features a signal for a non-sequential residue with a carbon frequency nearing that of the target C^α . The element-wise multiplication of equation (20b) consists in multiplying the amplitudes of signals at every point on the abscissa of spectra shown in Figure 5C (top two panels) or 5D (top two panels) to obtain spectra representing the vectors $\vec{v}_*$ and $\vec{v}_x$ (bottom panels in 5C and 5D). The summation of equation (20a) is represented by the area for these spectra shown in bottom panels of Figures 5C and 5D, which leads to the amplitudes of the true and false correlations H_{il}^* and H_{il}^x , respectively. This analysis highlights that a correlation will appear in COSCOMs as long as two slices feature common regions where signal amplitudes grow. In order to eliminate or reduce the false positives, it is necessary to emphasize matching of the peak maxima of the covaried signals and down-weight the amplitude overlap in the case of mismatched peak-maxima. This can be achieved by taking the derivative of the covaried signal prior to calculation of covariance (Figures 5F and 5G).

In the derivatives of spectra, a signal changes sign when a maximum is reached in the original spectrum, which now corresponds to an inflexion point. When two signals have identical maxima in the original spectra their product is always positive in the spectra after differentiation and summation of their elements is constructive. Thus, covariance of two spectra after calculating the derivative leads to correlations for signals with matching maxima. In contrast, when maxima are offset in the original spectra, zero-crossings happen at different abscissa after calculating the derivative and sign changes do not coincide. Thus, both positive and negative areas emerge in the product spectra (bottom panel in Figure 5G), and covarying two spectra after differentiation reduces, eliminates, or changes the sign of correlations for signals with mismatched maxima.

The efficiency of artifact elimination depends on the amount of overlap between signals with mismatched maxima. Figure 6 depicts the amplitude of a COSCOM correlation, corresponding to H_{ij} in equation (19), as a function of frequency offsets between two C^α signals with Lorentzian line-shapes.³⁸ The critical frequency separation, Δf_{crit} , where artifacts vanish, can be estimated by calculating the zero of an analytical expression of H_{ij} . For Lorentzian signals, $\Delta f_{crit} = \frac{1}{\sqrt{3}}FWHM_{av}$, where $FWHM_{av}$ is the average of the two peaks' full widths at half-max. For Gaussian signals, $\Delta f_{crit} = \frac{1}{2\pi}\sqrt{\sigma^2 + \sigma_x^2}$, where σ and σ_x depict the line-widths of the two signals compared. For both line-shapes, the amplitude of the correlation in the covariance spectrum decreases as the frequency separation increases, vanishes at the critical frequency, and becomes negative for larger values. Negative residual artifact correlations can be purged by extracting only the positive amplitudes of the correlation map. A word of caution is needed here. In Figure 6, the desired signal appears positive and residual artifacts above Δf_{crit} appear negative because the two input spectra leading to the covariance map have the same phase. If the input spectra have opposite phases, the signs are inverted in Figure 6. It is critical to know that the scripts used to calculate covariance maps extract positive amplitudes by default. If users supply spectra with opposite phases the covariance maps will only contain residual artifacts. With proper phasing, artifacts are actively suppressed for offsets larger than Δf_{crit} while they are reduced for smaller offsets. Consequently, calculating the derivatives of spectra becomes particularly advantageous when applied to high-resolution spectra, e.g. obtained through nonuniform sampling.

In this section, we have treated spectral comparisons as 2D or 1D covariance analyses. However, in practice, the correlation map is obtained using the GIC framework. In essence, the $[H][CA]$ and $[H][CA_s]$ planes are indexed as stacked spectra described in equation (10), and covariance calculations (including square-rooting) are carried out on the GIC matrix before extracting the relevant regions to create the 4D array $[H][N][H_s][N_s]$. A detailed description of the MATLAB script used to calculate the covariance is presented elsewhere, together with practical considerations for obtaining reliable maps.⁴⁰

The pairing of information provided by C^α , C^β , and C' is obtained by a simple mathematical operation that further eliminates artifacts. We have seen in the Introduction and Figure 1 that true carbon frequency degeneracies abound in larger proteins, and data from multiple pairs

of intra and sequential spectra must be considered for identifying sequential residues. In COSCOMs, these degeneracies lead to (H_r, N_r, H_x, N_x) correlations that cannot be distinguished from the true $(H_r, N_r, H_{r+1}, N_{r+1})$ correlations. However, in general, residues feature different accidental frequency degeneracies for different carbon nuclei. That is, accidental correlations (H_r, N_r, H_x, N_x) will appear at different indexes l and m for arrays $[H][N][H_s][N_s](i, j, l, m)$ obtained through covariance of $[CA]$ and $[CA_s]$, $[CB]$ and $[CB_s]$, or $[C']$ and $[C'_s]$ dimensions. Thus, an element-wise multiplication of these arrays selects for the true $(H_r, N_r, H_{r+1}, N_{r+1})$ correlation, while other correlations vanish as they are multiplied by spectral noise. We will see that the same advantage can be exploited during side-chain resonance assignments, and an example is shown in Figure 8. This consideration also applies to residual, artifact correlations stemming from carbon frequencies with offsets below Δf_{crit} . In the end, the combination of differentiating carbon dimensions and combining COSCOMs provides an array $[H][N][H_s][N_s]$ that faithfully reports on the combined information conveyed by HNCA, HN(CO)CA, HN(CA)CB, HN(COCA)CB, HN(CA)CO and HNCO.

COSCOMs allow for facile and reliable sequential resonance assignment. The 4D array $[H][N][H_s][N_s]$ can be described as a 2D array of 2D arrays. Here, to every correlation in an HSQC-like spectrum spanned by $[H][N]$ corresponds another HSQC spanned by $[H_s][N_s]$. The latter only features correlations that reflect simultaneous matching of carbon signals in all spectra used to generate the array. When C^α , C^β , and C' are combined, most $[H_s][N_s]$ planes feature a single correlation and the assignment is immediate. Here, the correlation map bypasses errors in peak picking and clearly identifies resonances for which additional data is necessary. It is not uncommon for residues to not display NMR signals, for example because of line-broadening or because a deuterated amide group exchanges too slowly with water to lead to a detected amide proton signal. In such cases, a considerable amount of time can be spent looking for resonances or verifying peak picking. In COSCOMs, the $[H_s][N_s]$ plane of the preceding residue will be devoid of correlations or display residual artifact correlations of low amplitude, clearly revealing that the signals of this residue are not detected in the original spectra. The advantages of the method are however best illustrated in Figure 1. In this example, amide proton and nitrogen frequencies are degenerate for two residues viz., G48 and T52, and carbon signals can only be paired by trial and error. In contrast, with COSCOMs the correlations for both sequential residues are readily obtained by visualization of the $[H_s][N_s]$ plane at the common (H_r, N_r) coordinate (Figure 7). The carbon chemical shifts are easily assigned by synchronizing original spectra with the newly found (H_{r+1}, N_{r+1}) coordinates and comparing carbon dimensions with those of intra spectra at (H_r, N_r) . Indeed, it is recommended to always synchronize the original data and inspect it simultaneously with $[H_s][N_s]$ planes as a number of scenarios can mitigate successful elimination of artifacts. For example, intra spectra also contain sequential correlations, and they may correlate with other sequential correlations in sequential spectra. Further, residual artifact correlations involving strong signals may display stronger amplitudes than that of a true sequential correlation with a residue displaying weak NMR signals. As a user will necessarily click on a signal to tabulate its assignment, synchronizing original spectra provides a safety net for identifying surviving artifacts without any loss in time. It is also possible to score correlations through more involved procedures,^{29,41,42} and these scores may be displayed through peak lists mapping the 4D arrays. In general, since the procedure

relies solely on mathematical operations, the features of input spectra propagate to the correlation maps. For example, the resolution and signal to noise of the covariance maps are limited by those of the input spectra, with the notable exception of relative signal amplitudes. Indeed, the many multiplications used to create novel correlations magnify the dynamic range observed in separate input spectra, and in 4D arrays the contour levels must be adjusted for each plane visualized. On the other hand, multiplying spectra may help overcome low signal-to-noise to some extent. Indeed, we observed that weak correlations in an input spectrum can be rescued when they are multiplied with stronger correlations in other spectra.³⁸ In the end, through covariance analysis and element-wise multiplication, user intervention is only required at the outset, once the information of all spectra has been combined to identify candidate sequential residues.

Covariance NMR for obtaining methyl assignments in large proteins:

Simultaneous analysis of multiple sets of spectra is also critical for accurate assignment of methyl signals of isoleucine, leucine, and valine residues (ILV residues) in large proteins. Here, methyl resonances are assigned by pairing them with assigned amide anchors through comparison of C^α and C^β signals. The 3D HMCMCBCA and HMCM(CG)CBCA experiments (HMCM spectra)²⁰ correlate the methyl resonances of a residue with those of both their C^α and C^β carbon nuclei to provide spectra of the form [HM][CM][CACB], where [CACB] contains signals of both C^α and C^β carbons. These correlations can be compared with those of [H][N][CA] and [H][N][CB] spectra to identify common C^α and C^β signals, thus transferring assignments of amide moieties to methyl groups. This time-consuming procedure consists in comparing strips containing C^α and C^β signals for all (HM,CM) and (H,N) correlations, which can be formulated instead as a covariance analysis between [CACB] and [CA] or [CB] dimensions (Figure 8A).⁴³

$$[H][N][HM][CM](i, j, l, m)_{CA} = \sum_{k=1}^K [H][N][CA](i, j, k)[HM][CM][CACB](l, m, k), (21a)$$

$$[H][N][HM][CM](i, j, l, m)_{CB} = \sum_{k=1}^K [H][N][CB](i, j, k)[HM][CM][CACB](l, m, k).$$

(21b)

Element-wise multiplication of both [H][N][HM][CM] arrays eliminates artifacts due to C^α or C^β chemical shift degeneracy (Figure 8B). Unfortunately, the experiments correlating methyls with carbonyl carbons were too insensitive for practical implementation and the maps must rely on C^α and C^β resonances only. Further, [CACB] dimensions contain signals of both C^α and C^β carbons, as well as multiple-quantum artifacts, which can all correlate with resonances in [CA] and [CB] dimensions of amide spectra. Thus, many artifacts appear in [HM][CM] planes if entire spectra are used during calculation of covariance maps. However, here, the objective is to assign methyl resonances of specific residue types viz., isoleucines, valines, and leucines. Hence, it is possible to calculate residue-specific

correlation maps by extracting spectral regions specific to each amino acid from the original spectra. The procedure dramatically reduces the number of artifacts as only specific, targeted regions of the carbon dimensions are covaried. Figure 8B shows that, when combined with derivative and element-wise multiplication, these residue-specific [H][N][HM][CM] arrays reveal the methyl correlations of an assigned amide anchor by simple visual inspection. Thus, the assignment procedure simply consists in clicking on (H,N) coordinates and inspecting [HM][CM] planes (Figure 8A).

CONCLUSIONS:

Many advances were necessary to permit applications of covariance NMR to protein resonance assignment. Notably, indirect covariance, and the general indirect covariance formalism, enabled for correlating two different spectra featuring a common dimension. Matrix square-rooting had to be integrated into this formalism to prevent relay artifacts. Further improvements were spurred by applications to large proteins. Thus, spectral derivative of covaried dimensions and element-wise multiplication of correlation maps obtained through different NMR spectra were both needed to provide reliable correlation maps in presence of signal overlap. With these advances, it became possible to generate arrays that directly correlate resonances of interest. Thus, instead of assigning sequential residues by inspection of six carbon dimensions, a single correlation capturing this information identifies sequential residues. Similarly, methyl resonances can be assigned from amide groups by simple visual inspection of a residue-specific correlation map.

Although covariance maps facilitate resonance assignments, they are nevertheless meant to be integrated to existing procedures rather than replace them. Covariance maps readily identify candidate sequential residues from the *combined* information of all spectra, thus bypassing limitations of preliminary peak picking. However, the original data enable pairing carbon resonances to the residues assigned whilst providing a checkpoint for artifact correlations, e.g. due to sequential correlations in intra spectra. Similarly, covariance NMR is not intended to replace automated assignment procedures but to ensure that the data used for automated assignment is reliable. Automated assignment procedures rely on correct identification of signals belonging to sequential residues before matching their tentative residue types to sequences of residues in the protein primary structure. Without covariance maps, the protocols require peak picking of 3D spectra to find sequential residues and are necessarily vulnerable to the errors described in our introduction. Covariance maps accurately identify resonances of sequential residues so that automated resonance assignments match these sets of resonances with the proper sequence fragment.

It is worth emphasizing that the covariance method is not subject to experimental constraints and is applicable to both nonuniformly and uniformly acquired spectra. Indeed, covariance maps can be calculated as long as subsumed dimensions are referenced and feature the same number of points per ppm. Similarly, covariance maps can be combined through element-wise multiplication when common dimensions feature the same number of points per ppm. These conditions are easily satisfied through proper zero-filling of relevant dimensions, which can be performed after data acquisition. Clearly, the order of the dimensions must be consistent, e.g. [H][N][CA] is combined with [H_s][N_s][CA_s] and not [N_s][CA_s][H_s], which

is easily resolved through data transposition using processing software packages such as nmrPipe⁴⁴. In conclusion, covariance maps offer immediate support in resonance assignments by combining information otherwise provided through separate spectra, and only require consistent processing of original spectra and running a matlab script. We provide the script on our website (<http://frueh.med.jhmi.edu/software-downloads/>) together with examples, and its usage is described in detail elsewhere.⁴⁰ Covariance maps are calculated in minutes, and visual inspection rapidly provides new assignments or reveals incorrect assignments. Thus, when six months were spent to assign a 37 kDa protein, 70% of its backbone resonances were assigned in two weeks when covariance maps were included.³⁹ The method can be readily used at the onset of resonance assignment or to rescue ongoing projects without acquiring new data, and it should provide a useful tool for resonance assignments in crowded NMR spectra.

ACKNOWLEDGEMENTS:

We thank Bradley Harden, Indrani Pal and Subrata Mishra for helpful discussions. Research in the Frueh Lab is supported by NIH grant R01 GM104257 and our equipment benefitted from grant S10 RR029191.

Biographies

Aswani K. Kancherla completed his undergraduate studies in Biotechnology in 2006 at the Vellore Institute of Technology, Tamil Nadu, India. He pursued his graduate studies in the laboratory of Siddhartha P. Sarma, at the Indian Institute of Science, Bangalore, and obtained a Ph.D. in 2016 for his NMR studies on the structures and dynamics of conotoxins as well as for his studies of a nuclease domain from *Helicobacter pylori*. Since July 2016, he has been a post-doctoral researcher in the laboratory of Dominique Frueh at the Johns Hopkins University School of Medicine, Baltimore, USA. He is currently studying the molecular basis for communication between a nonribosomal peptide synthetase cyclization domain and its partner carrier protein by probing an active enzymatic system by NMR.

Dominique P. Frueh obtained his Diplome de Chimie from the Université de Lausanne in 1997 (faculty prize 1995, math excellence 1992). He then joined the laboratory of Geoffrey Bodenhausen to develop novel solution NMR methods with a focus on protein dynamics, and obtained his PhD from the EPFL in 2002. He was a post-doctoral fellow with a Swiss National Science Foundation fellowship (2002–2003) and went to Harvard Medical School, Boston, MA to work with Gerhard Wagner and Christopher T Walsh, and was later promoted as an Instructor and Research Associate (2007–2010). In 2010, he joined the department of Biophysics and Biophysical Chemistry at the Johns Hopkins School of Medicine, Baltimore, MD, and was promoted to associate professor in 2016. His laboratory studies the molecular mechanisms of nonribosomal peptide synthetases and develops NMR methods for studies of large and/or dynamic proteins.

REFERENCES:

1. Williamson MP. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc.* 2013;73:1–16. [PubMed: 23962882]
2. Eisenmesser EZ, Millet O, Labeikovsky W, et al. Intrinsic dynamics of an enzyme underlies catalysis. *Nature.* 2005;438(7064):117–121. [PubMed: 16267559]

3. Goodrich AC, Harden BJ, Frueh DP. Solution Structure of a Nonribosomal Peptide Synthetase Carrier Protein Loaded with Its Substrate Reveals Transient, Well-Defined Contacts. *J Am Chem Soc.* 2015;137(37):12100–12109. [PubMed: 26334259]
4. Selenko P, Frueh DP, Elsaesser SJ, Haas W, Gygi SP, Wagner G. In situ observation of protein phosphorylation by high-resolution NMR spectroscopy. *Nat Struct Mol Biol.* 2008;15(3):321–329. [PubMed: 18297086]
5. Theillet FX, Binolfi A, Bekei B, et al. Structural disorder of monomeric α -synuclein persists in mammalian cells. *Nature.* 2016;530(7588):45–50. [PubMed: 26808899]
6. Bodenhausen G, Ruben DJ. Natural abundance Nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem Phys Lett.* 1980;69(1):185–189.
7. Jeener J, Meier BH, Bachmann P, Ernst RR. Investigation of exchange processes by two-dimensional NMR spectroscopy. *J Chem Phys.* 1979;71(11):4546–4553.
8. Kay LE, Ikura M, Tschudin R, Bax A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J Magn Reson.* 1990;89(3):496–514.
9. Grzesiek S, Bax A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J Magn Reson.* 1992;96(2):432–440.
10. Ikura M, Kay LE, Bax A. A novel approach for sequential assignment of ^1H , ^{13}C , and ^{15}N spectra of larger proteins: Heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry.* 1990;29(19):4659–4667. [PubMed: 2372549]
11. Bax A, Ikura M. An efficient 3D NMR technique for correlating the proton and ^{15}N backbone amide resonances with the α -carbon of the preceding residue in uniformly $^{15}\text{N}/^{13}\text{C}$ enriched proteins. *J Biomol NMR.* 1991;1(1):99–104. [PubMed: 1668719]
12. Wittekind M, Mueller L. HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the α - and β -carbon resonances in proteins. *J Magn Reson Ser B.* 1993;101(2):201–205.
13. Clubb RT, Thanabal V, Wagner G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue $^1\text{H}^{\text{N}}$, ^{15}N , and $^{13}\text{C}'$ chemical shifts in $^{15}\text{N}/^{13}\text{C}$ -labelled proteins. *J Magn Reson.* 1992;97(1):213–217.
14. Frueh DP. Practical aspects of NMR signal assignment in larger and challenging proteins. *Prog Nucl Magn Reson Spectrosc.* 2014;78:47–75. [PubMed: 24534088]
15. Keller R. The computer aided resonance assignment tutorial. 2004.
16. Goddard TD, Kneller DG. Sparky 3. San Fr Univ Calif.
17. Vranken WF, Boucher W, Stevens TJ, et al. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins Struct Funct Bioinforma.* 2005;59(4):687–696.
18. Grzesiek S, Anglister J, Bax A. Correlation of backbone amide and aliphatic side-chain resonances in $^{13}\text{C}/^{15}\text{N}$ -enriched proteins by isotropic mixing of ^{13}C magnetization. *J Magn Reson Ser B.* 1993;101(1):114–119.
19. Kay LE, Xu GY, Singer AU, Muhandiram DR, Formankay JD. A Gradient-Enhanced HCCH-TOCSY Experiment for Recording Side-Chain ^1H and ^{13}C Correlations in H_2O Samples of Proteins. *J Magn Reson Ser B.* 1993;101(3):333–337.
20. Tugarinov V, Kay LE. Ile, Leu, and Val Methyl Assignments of the 723-Residue Malate Synthase G Using a New Labeling Strategy and Novel NMR Methods. *J Am Chem Soc.* 2003;125(45):13868–13878. [PubMed: 14599227]
21. Snyder DA, Brüschweiler R. Chapter 10 Multi-dimensional Spin Correlations by Covariance NMR. *Mod NMR Approaches to Struct Elucidation Nat Prod.* 2015;1:244–258.
22. Brüschweiler R, Zhang F. Covariance nuclear magnetic resonance spectroscopy. *J Chem Phys.* 2004;120(11):5253–5260. [PubMed: 15267396]
23. Brüschweiler R, Brüschweiler R. Theory of covariance nuclear magnetic resonance spectroscopy. *J Chem Phys.* 2004;121(1):409–414. [PubMed: 15260561]
24. Lipkus AH, Nieman RA, Munk ME. A manipulation of two-dimensional NMR spectra based on graph theory. *J Magn Reson - Ser A.* 1993;102(1):24–28.
25. Zhang F, Brüschweiler R. Indirect covariance NMR spectroscopy. *J Am Chem Soc.* 2004;126(41):13180–13181. [PubMed: 15479045]

26. Fesik SW, Eaton HL, Olejniczak ET, Zuiderweg ERP, McIntosh LP, Dahlquist FW. 2D and 3D NMR Spectroscopy Employing ^{13}C -, ^{13}C Magnetization Transfer by Isotropic Mixing. Spin System Identification in Large Proteins. *J Am Chem Soc.* 1990;112(2):886–888.
27. Blinov KA, Larin NI, Williams AJ, Zell M, Martin GE. Long-range carbon-carbon connectivity via unsymmetrical indirect covariance processing of HSQC and HMBC NMR data. *Magn Reson Chem.* 2006;44(2):107–109. [PubMed: 16392113]
28. Bax A, Summers MF. ^1H and ^{13}C Assignments from Sensitivity-Enhanced Detection of Heteronuclear Multiple-Bond Connectivity by 2D Multiple Quantum NMR. *J Am Chem Soc.* 1986;108(8):2093–2094.
29. Snyder DA, Bruschweiler R. Generalized Indirect Covariance NMR Formalism for Establishment of Multidimensional Spin Correlations. *J Phys Chem A.* 2009;113:12898–12903. [PubMed: 19810742]
30. Martin GE, Hilton BD, Irish PA, Blinov KA, Williams AJ. Using unsymmetrical indirect covariance processing to calculate GHSQC-COSY spectra. *J Nat Prod.* 2007;70(9):1393–1396. [PubMed: 17691750]
31. Kup e E, Freeman R. Natural-abundance ^{15}N - ^{13}C correlation spectra of vitamin B-12. *Magn Reson Chem.* 2007;45(2):103–105. [PubMed: 17201010]
32. Martin GE, Irish PA, Hilton BD, Blinov KA, Williams AJ. Utilizing unsymmetrical indirect covariance processing to define ^{15}N - ^{13}C connectivity networks. *Magn Reson Chem.* 2007;45(8):624–627. [PubMed: 17563910]
33. Martin GE, Hilton BD, Irish PA, Blinov KA, Williams AJ. Application of unsymmetrical indirect covariance NMR methods to the computation of the $^{13}\text{C} \leftrightarrow ^{15}\text{N}$ HSQC-IMPEACH and $^{13}\text{C} \leftrightarrow ^{15}\text{N}$ HMBC-IMPEACH correlation spectra. *Magn Reson Chem.* 2007;45(10):883–888. [PubMed: 17729230]
34. Kup e E, Freeman R. Hyperdimensional NMR spectroscopy. *J Am Chem Soc.* 2006;128(18):6020–6021. [PubMed: 16669655]
35. Chen K, Delaglio F, Tjandra N. A practical implementation of cross-spectrum in protein backbone resonance assignment. *J Magn Reson.* 2010;203(2):208–212. [PubMed: 20053573]
36. Blinov KA, Larin NI, Kvasha MP, Moser A, Williams AJ, Martin GE. Analysis and elimination of artifacts in indirect covariance NMR spectra via unsymmetrical processing. *Magn Reson Chem.* 2005;43(12):999–1007. [PubMed: 16144032]
37. Trbovic N, Smirnov S, Zhang F, Bruschweiler R. Covariance NMR spectroscopy by singular value decomposition. 2004;171:277–283.
38. Harden BJ, Mishra SH, Frueh DP. Effortless assignment with 4D covariance sequential correlation maps. *J Magn Reson.* 2015;260:83–88. [PubMed: 26432397]
39. Harden BJ, Nichols SR, Frueh DP. Facilitated assignment of large protein NMR signals with covariance sequential spectra using spectral derivatives. *J Am Chem Soc.* 2014;136(38):13106–13109. [PubMed: 25226241]
40. Harden BJ, Frueh DP. Chapter 16 Covariance NMR Processing and Analysis for Protein. *Methods Mol Biol.* 2018;1688:353–373. [PubMed: 29151217]
41. Wei Q, Chen J, Mi J, Zhang J, Ruan K, Wu J. NMR Backbone Assignment of Large Proteins by Using ^{13}Ca -Only Triple-Resonance Experiments. *Chem - A Eur J.* 2016;22(28):9556–9564.
42. Snyder DA, Ghosh A, Zhang F, Szyperski T, Bruschweiler R. Z-matrix formalism for quantitative noise assessment of covariance nuclear magnetic resonance spectra. *J Chem Phys.* 2008;129(10).
43. Mishra SH, Frueh DP. Assignment of methyl NMR resonances of a 52 kDa protein with residue-specific 4D correlation maps. *J Biomol NMR.* 2015;62(3):281–290. [PubMed: 25953312]
44. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* 1995;6(3):277–293. [PubMed: 8520220]

FIGURE 1.

Sequence specific backbone resonance assignments through traditional strip matching. (A) $[H][C]$ strips are extracted for each (H,N) correlation from the three complementary pairs of 3D spectra HNCA and HN(CO)CA (top panel); HN(CA)CO and HNCO (middle panel); and HN(CA)CB and HN(COCA)CB (bottom panel), respectively. The correlations in “intra” spectra are shown in red while the correlations in “sequential” spectra are shown in blue. (B) Nuclei that show correlations in “intra” and “sequential” spectra are highlighted on a polypeptide fragment using red and blue squares, respectively. (C) Sequential residues are identified by comparing strips from sequential spectra (blue) with that of a target in the intra spectra (red). In the example shown, two residues (G48 and T52 from a 52 kDa nonribosomal peptide synthetase domain Cyl) overlap in $[H][N]$ dimensions and the C^α , C' and, C^β signals must be tentatively paired to explore the solutions. In addition, each carbon signal has frequencies matching those of several other residues (selected examples are shown) making the assignment process further complicated. Only proper pairing identifies the signals of the sequential residues (shown in bold frame with gray background). Figures 1A and B have been adapted from Frueh (2014)¹⁴ with permission from Elsevier.

FIGURE 2.

Representation of NMR spectra as matrices. (A) A one-dimensional NMR spectrum with M data points is represented as a one-dimensional vector with M elements, where each element denotes signal amplitude. (B) A two-dimensional NMR spectrum with $(M \times N)$ data points is represented as an $(M \times N)$ matrix with elements carrying signal amplitudes.

FIGURE 3.

Indirect covariance NMR applied to an HC-HSQC-TOCSY spectrum to obtain a $[\text{C}_\text{t}][\text{C}_\text{t}]$ -TOCSY correlation map. (A) HC-HSQC-TOCSY spectrum of a mixture of glutamine, lysine, and valine at 130 mM concentration. The proton correlations that belong to valine are emphasized by dashed lines at the corresponding carbon frequencies. (B) CC-TOCSY spectrum obtained by indirect covariance NMR using equation (5). The carbon correlations that are obtained by subsuming the proton dimension in (A) are highlighted by dashed lines for valine. Figure reproduced from Zhang and Bruschweiler (2004)²⁵ with permission from the American Chemical Society.

FIGURE 4.

Suppression of artifacts in covariance spectra by matrix square-rooting. Simulated HMBC (A), HH-TOCSY (B), and HMBC-TOCSY obtained by covarying the proton dimension of HMBC and TOCSY without (C) and with (D) matrix square-rooting. The simulation reports on a spin system with three connected ^{13}C - ^1H pairs (x, y, and z, shown in green in (E)) and a spin system with two connected ^{13}C - ^1H pairs (u, v, shown in blue in (E)). The degeneracy in chemical shift for the protons H_Y and H_U leads to false positive signals (red) in the HMBC-TOCSY covariance spectrum without square-rooting (C). Application of the matrix square-root eliminates false positives or reduces their intensity (D). Figure adapted from Snyder and Bruschweiler (2009)²⁹ with permission from the American Chemical Society.

FIGURE 5.

Covariance NMR and spectral derivatives to obtain sequence specific assignments with minimal artifacts. For all nitrogen indices, $[H][C]$ planes from the “intra” spectrum HNCA (A) and $[H_s][C_s]$ planes from the “sequential” spectrum HN(CO)CA (B) are covaried along the carbon dimension to obtain $[H][H_s]$ correlation maps (E). The correlation of the true sequential residue is shown in blue (*) while a signal with accidental partial overlap is shown in red (x). Element-wise multiplication of vectors along the carbon dimensions of HNCA (black, top-panel in (C) and (D)) and HN(CO)CA leads to vectors $\vec{v}_*$ and $\vec{v}_x$ reporting on covariance upon element summation. The vectors of both true (C) and accidental (D) correlations display a significant area leading to correlations in $[H][H_s]$ (E). With spectral derivative along the carbon dimension, the area of the vector $\vec{v}_*$ reporting on the true correlation remains positive (F, bottom panel) while that of the vector $\vec{v}_x$ reporting on the

accidental correlation changes sign (G, bottom panel), and only the true correlation appears in $[H][H_s]$ (H). Reporting planes for all nitrogen indices yields a 4D array $[N][H][N_s][H_s]$ displaying candidate sequential correlations for every amide moiety. Figure adapted from Harden et. al., (2014)³⁹ (<http://pubs.acs.org/doi/abs/10.1021/ja5058407>) with permission from the American Chemical Society. Further permissions related to this material should be directed to ACS.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

FIGURE 6.

Near-degeneracy artifacts in covariance maps. (A) covariance peak amplitude as a function of frequency offset Δf for the maxima of the partially overlapped peaks. The amplitude of the covariance correlation decreases with increasing separation, becomes zero at the critical separation Δf_{crit} , and becomes negative for larger separations. (B) The inflection points of Lorentzian line-shapes align when they are separated by Δf_{crit} . Figure reproduced from Harden et. al., (2015)³⁸ with permission from Elsevier.

FIGURE 7.

4D COSCOMs of the nonribosomal peptide synthetase domain Cy1 for overlapping residues G48 and T52. An HN-TROSY spectrum of Cy1 is shown on the left. A 2D [$^1\text{H}_\text{S}$][$^1\text{N}_\text{S}$] plane at the (H, N) correlation of G48/T52 reveals the correlations of two sequential residues. The location of the intra-residue (H, N) correlation is shown using a red plus sign. The two unlabeled weak correlations result from residues with similar C^α , C^β , and C' frequencies. The conventional assignment process using strip matching for these two residues is shown in Figure 1.

FIGURE 8.

Covariance correlation maps for assigning methyl resonances. (A) Covariance is calculated both between [CA] dimensions of HNCA and HMCMCBCA and between [CB] of HN(CA)CB and HMCMCBCA, and element-wise multiplication leads to reliable 4D arrays. (B) Advantages of spectral derivative and element-wise multiplication when assigning the methyls of valine 315 in the nonribosomal peptide synthetase Cy1 domain. The red asterisks denote the methyl signals of V315. ○ denotes element-wise multiplication of maps obtained through [CA] (left) and [CB] (right) dimensions. Artefacts removed by element-wise multiplication are labeled with red circles. Top: calculation of 4D arrays without spectral derivatives in covaried [CA]/[CB] dimension. Bottom: covariance maps calculated after applying derivative in the covaried carbon dimensions. Artefacts removed by taking the spectral derivative are labeled with red diamonds. Combination of spectral derivative and element-wise multiplication significantly reduces the number and intensity of artifacts and emphasizes the true methyl correlations (bottom right). Figure adapted from Mishra and Frueh (2015)⁴³ with permission from Springer Nature.